吉田塁先生Project・東京財団政策研究所_2025年3月21日

先進的生成AIシステムとAI人材育成

須藤　修

＊中央大学国際情報学部　教授

＊中央大学ELSIセンター　所長

＊東京財団政策研究所 研究主幹

＊東京大学名誉教授

＊Member of the OECD Expert Group on AI Risk and Accountability

＊Member of the OECD Expert Group on AI Futures

＊内閣府「人間中心のAI社会原則会議」議長

＊総務省「AIネットワーク社会推進会議」議長

[1] LLM生成AI、マルチモーダルAI、MoE

Global AI Power Rankings: Stanford HAI Tool Ranks 36 Countries in AI, Nov. 21,2024 (42個の指標で評価)

[Source] Stanford HAI ed., Global AI Power Rankings: Stanford HAI Tool Ranks 36 Countries in AI, Nov. 21,2024

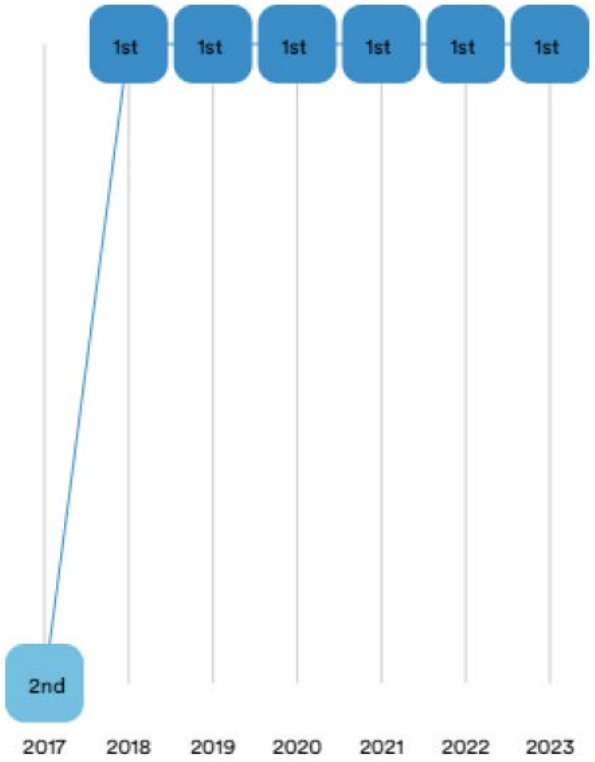
2023 AI Vibrancy Profile

United States

Overall Rank: **1st**

Absolute Metrics

Ranking Over Time



Top Ranking 2023

Two-Year Overview

	2023 Score	2023 Rank	2022 Score	2022 Rank
R&D	19.29	1/36	19.81	1/36
Responsible AI	5.71	1/36	5.71	1/36
Economy	13.55	1/36	13.88	1/36
Education	2.75	4/36	2.97	3/36
Diversity	1.01	27/36	0.87	26/36
Policy and Governance	6.84	6/36	10.48	1/36
Public Opinion	3.99	1/36	4.45	1/36
Infrastructure	16.91	1/36	12.47	2/36

[Source] Stanford HAI ed., Global AI Power Rankings: Stanford HAI Tool Ranks 36 Countries in AI, Nov. 21,2024

2023 AI Vibrancy Profile

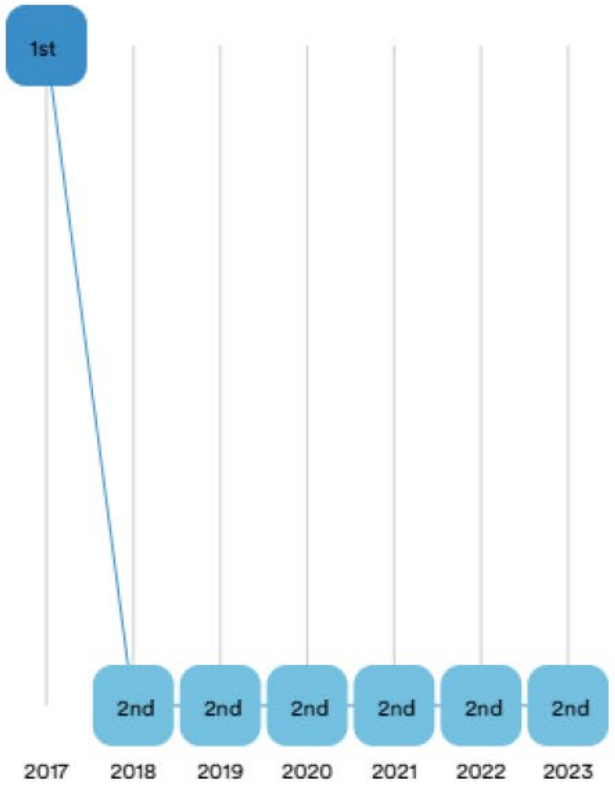
China

Overall Rank:

2nd

Absolute Metrics

Ranking Over Time



Top Ranking 2023

Two-Year Overview

	2023		2022	
	Score	Rank	Score	Rank
R&D	14.78	2/36	15.39	2/36
Responsible AI	1.96	2/36	1.75	2/36
Economy	6.19	2/36	7.90	2/36
Education	0.94	24/36	1.10	24/36
Diversity	1.08	13/36	0.91	13/36
Policy and Governance	4.40	33/36	4.41	32/36
Public Opinion	1.33	7/36	1.66	4/36
Infrastructure	9.49	2/36	13.32	1/36

[Source] Stanford HAI ed., Global AI Power Rankings: Stanford HAI Tool Ranks 36 Countries in AI, Nov. 21,2024

2023 AI Vibrancy Profile

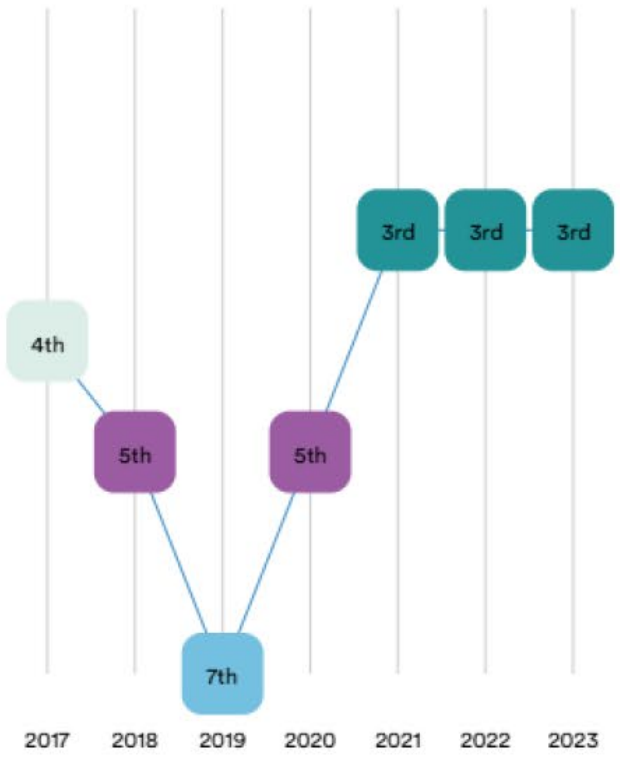
United Kingdom

Overall Rank:

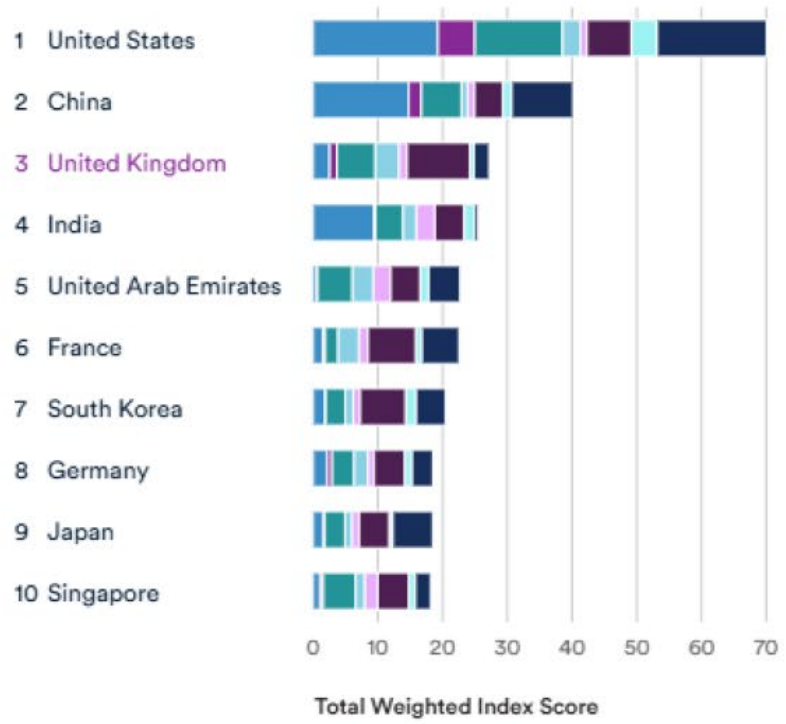
3rd

Absolute Metrics

Ranking Over Time



Top Ranking 2023



Two-Year Overview

	2023 Score	2023 Rank	2022 Score	2022 Rank
R&D	2.58	4/36	4.20	4/36
Responsible AI	1.11	3/36	1.22	3/36
Economy	5.92	4/36	7.03	3/36
Education	3.69	1/36	3.72	1/36
Diversity	1.25	8/36	1.16	8/36
Policy and Governance	9.67	1/36	7.16	3/36
Public Opinion	0.65	19/36	1.23	17/36
Infrastructure	2.35	21/36	1.65	23/36

[Source] Stanford HAI ed., Global AI Power Rankings: Stanford HAI Tool Ranks 36 Countries in AI, Nov. 21,2024

2023 AI Vibrancy Profile

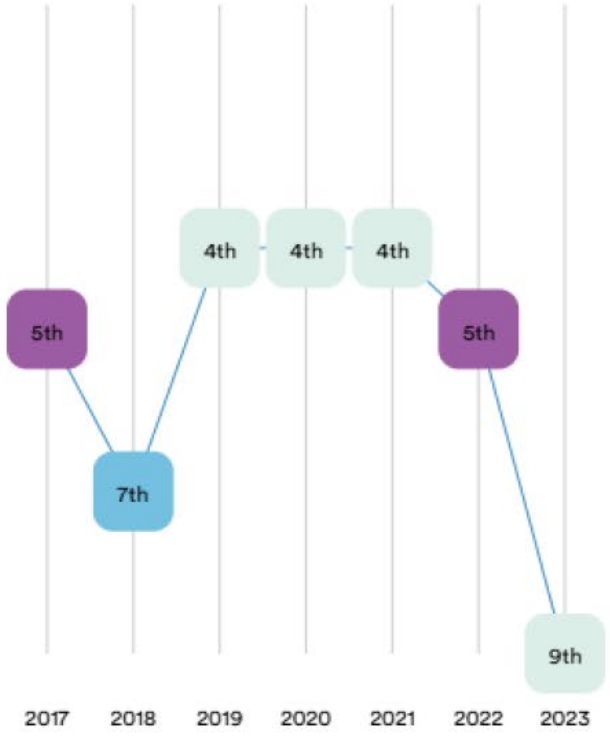
Japan

Overall Rank:

9th

Absolute Metrics

Ranking Over Time



Top Ranking 2023

Two-Year Overview

	2023		2022	
	Score	Rank	Score	Rank
R&D	1.61	8/36	1.74	8/36
Responsible AI	0.20	13/36	0.24	12/36
Economy	3.11	11/36	3.55	14/36
Education	1.11	23/36	1.46	17/36
Diversity	1.08	13/36	0.91	13/36
Policy and Governance	4.68	17/36	4.94	13/36
Public Opinion	0.54	24/36	0.97	27/36
Infrastructure	6.14	3/36	8.11	3/36

[Source] Stanford HAI ed., Global AI Power Rankings: Stanford HAI Tool Ranks 36 Countries in AI, Nov. 21,2024

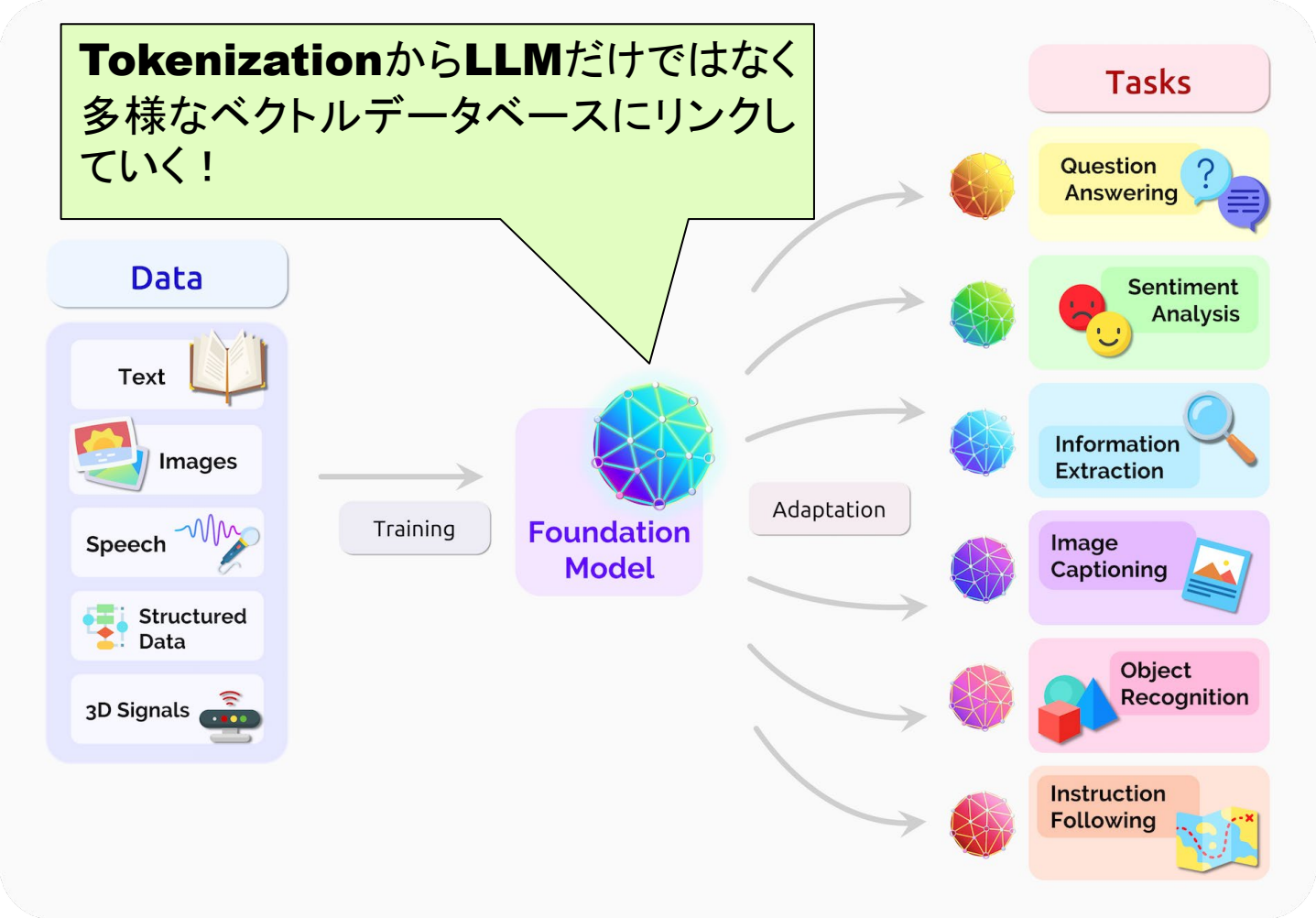
- ✓ いまや「AIの民主化」あるいは「AIの大衆化」と称される大きな変化が起こっている。

  高度なAIをコードではなく、自然言語で扱えるようになってきた。

■これから、LLM生成AIを基盤にしたマルチモーダルAIの研究開発、そして社会的普及が重要になる。

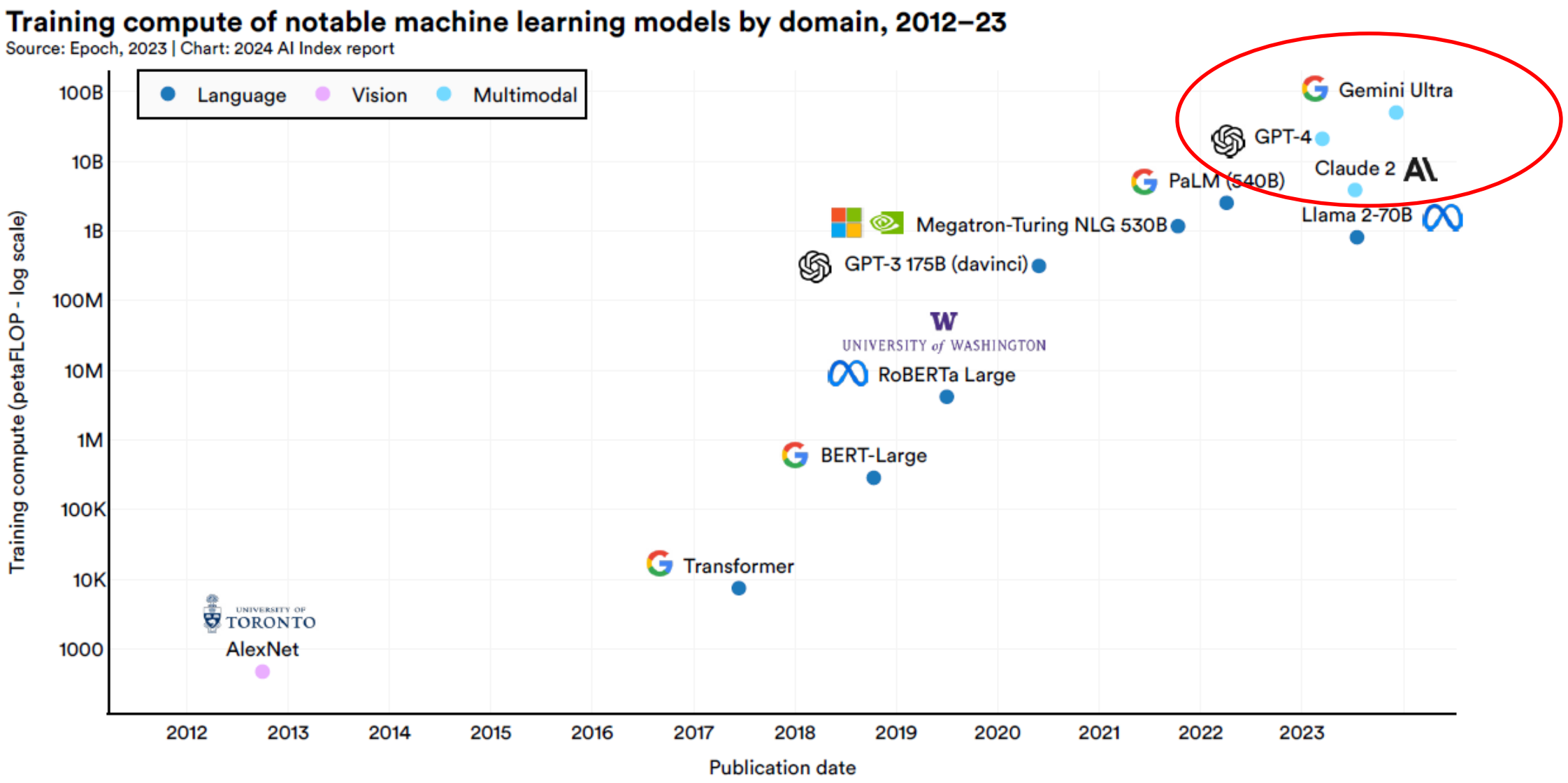
■これから、大学の教育の在り方、研究の在り方に多大な影響を与えることになる。

Towards AGI
Foundation Model : Beyond specific purpose AI ?



A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

[Source] Rishi Bommasani et.al. On the Opportunities and Risks of Foundation Models, *arXiv,* 2022



[Source] Epoch AI Impact Assessment

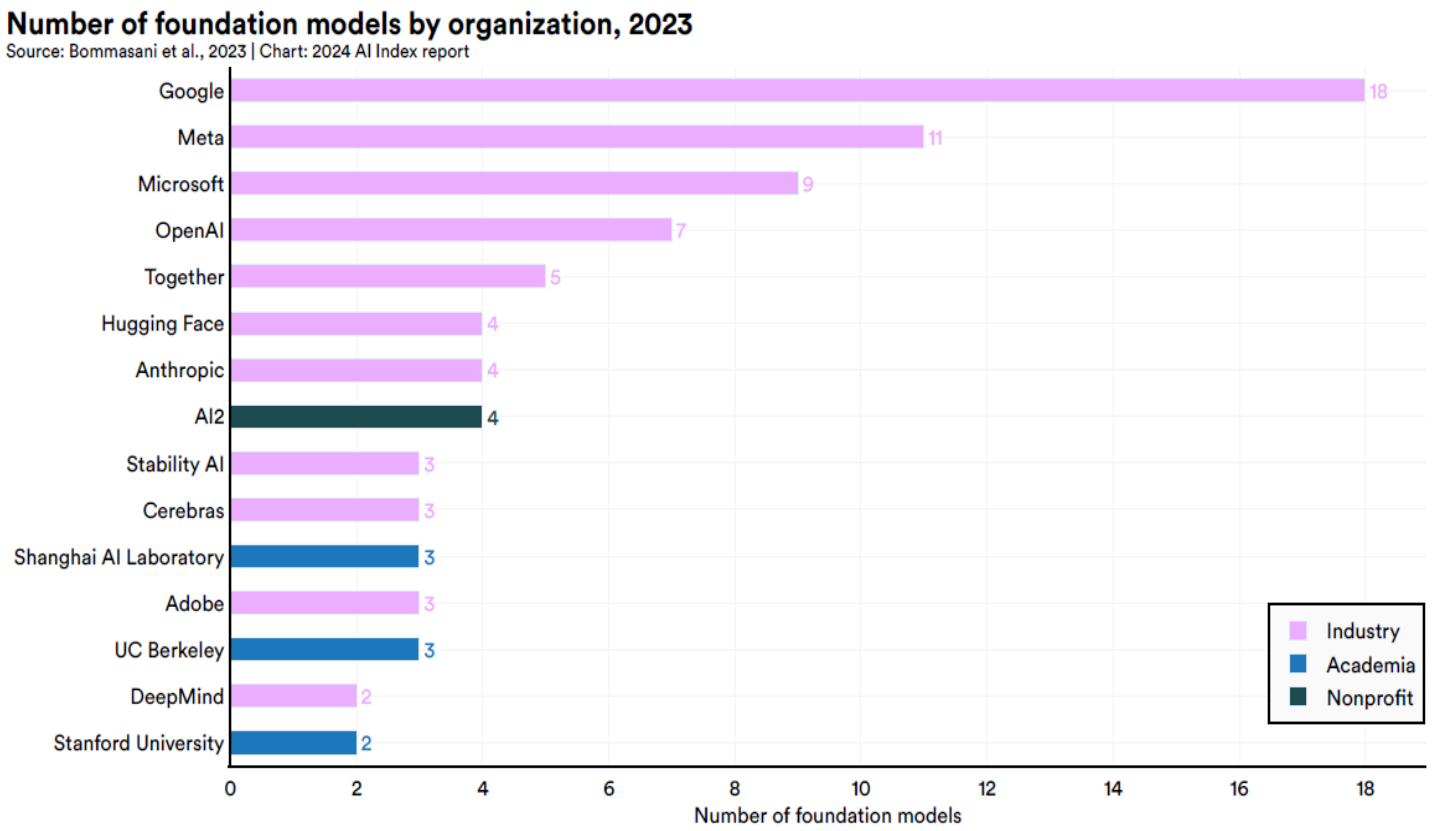
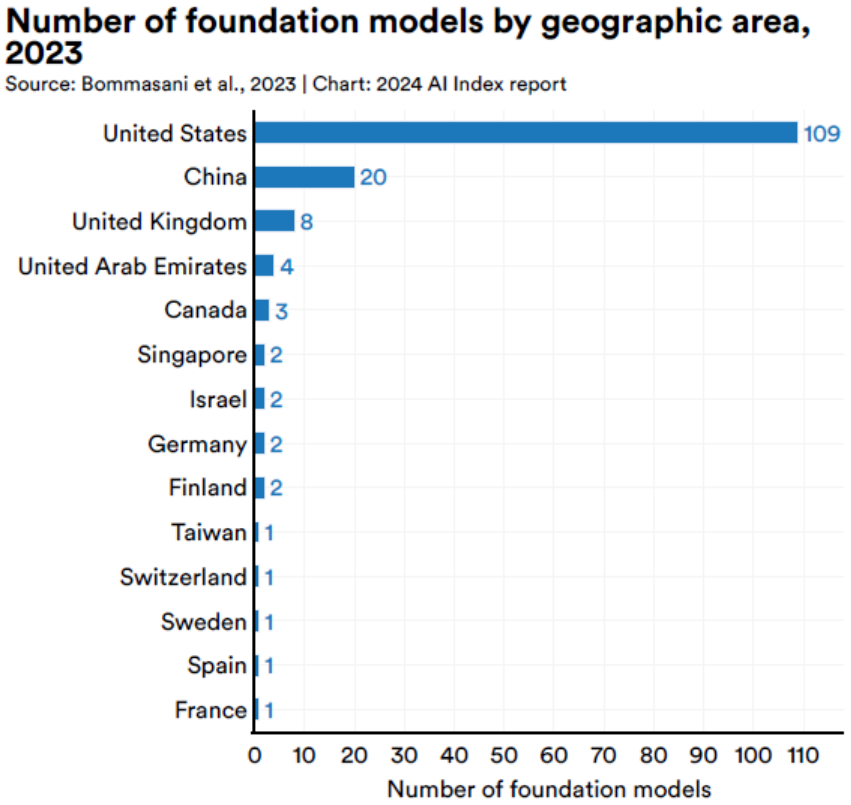
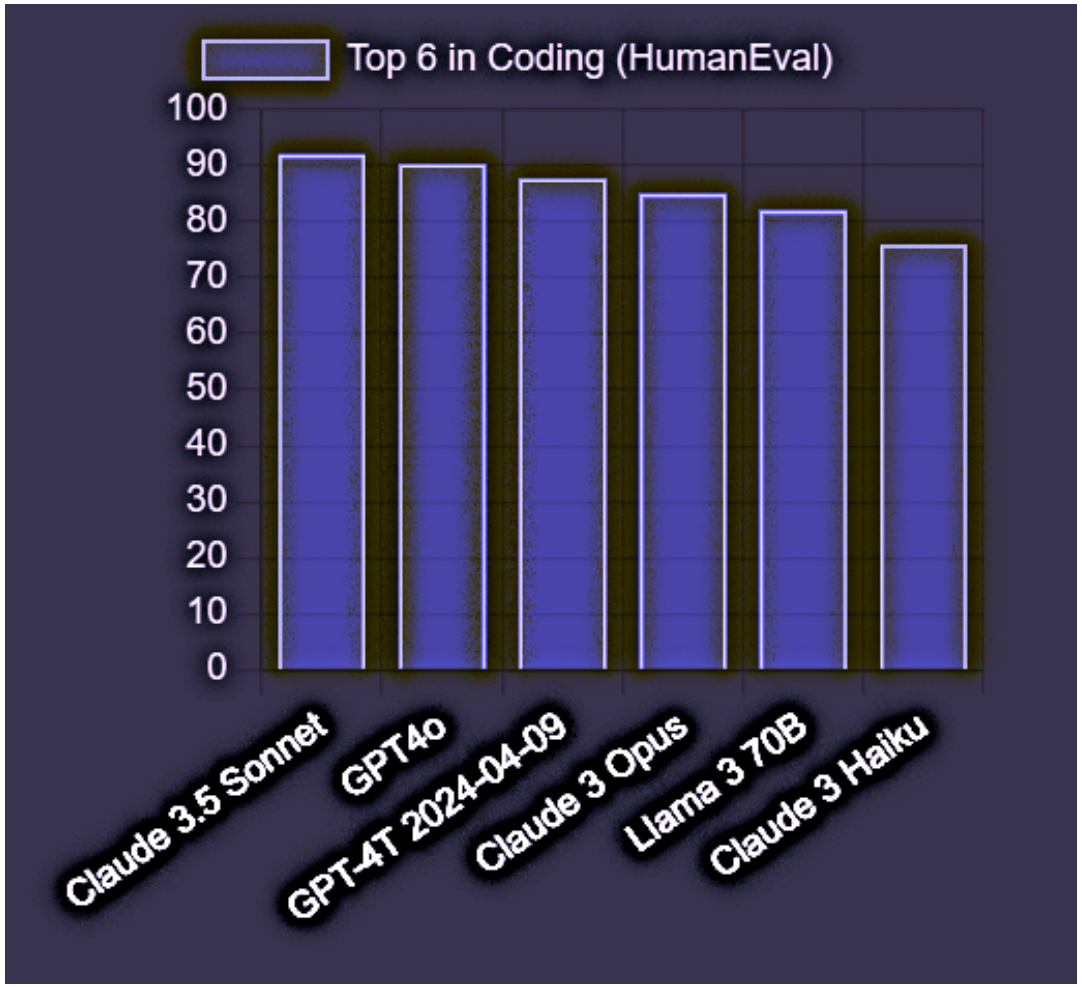


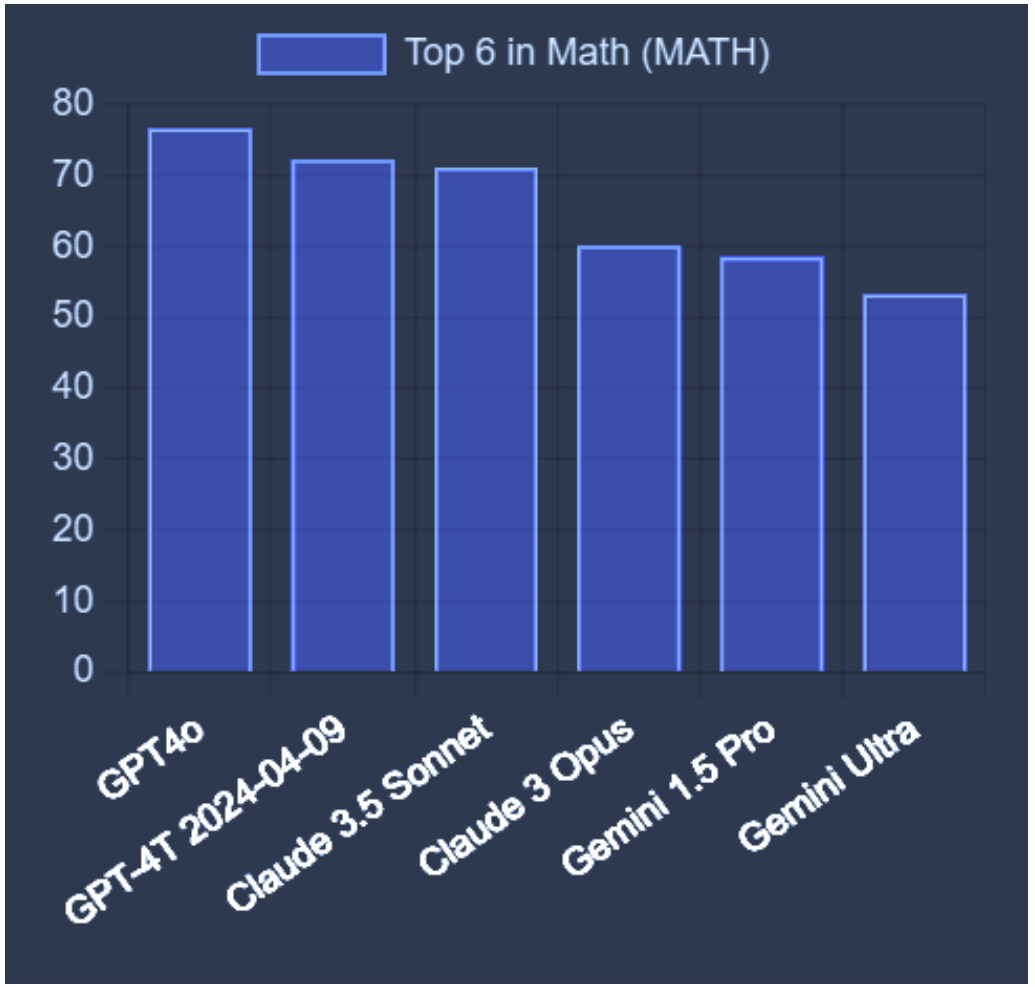
Figure 1.3.16



Source : Bommasani et.al., On the Opportunities and Risks of Foundation Models, *ArXiv*, 12 July 2023

マルチモーダルAIの性能検定





Source : LLM Safety Leaderboard 2024

DeepSeekのインパクト

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning DeepSeek-AI

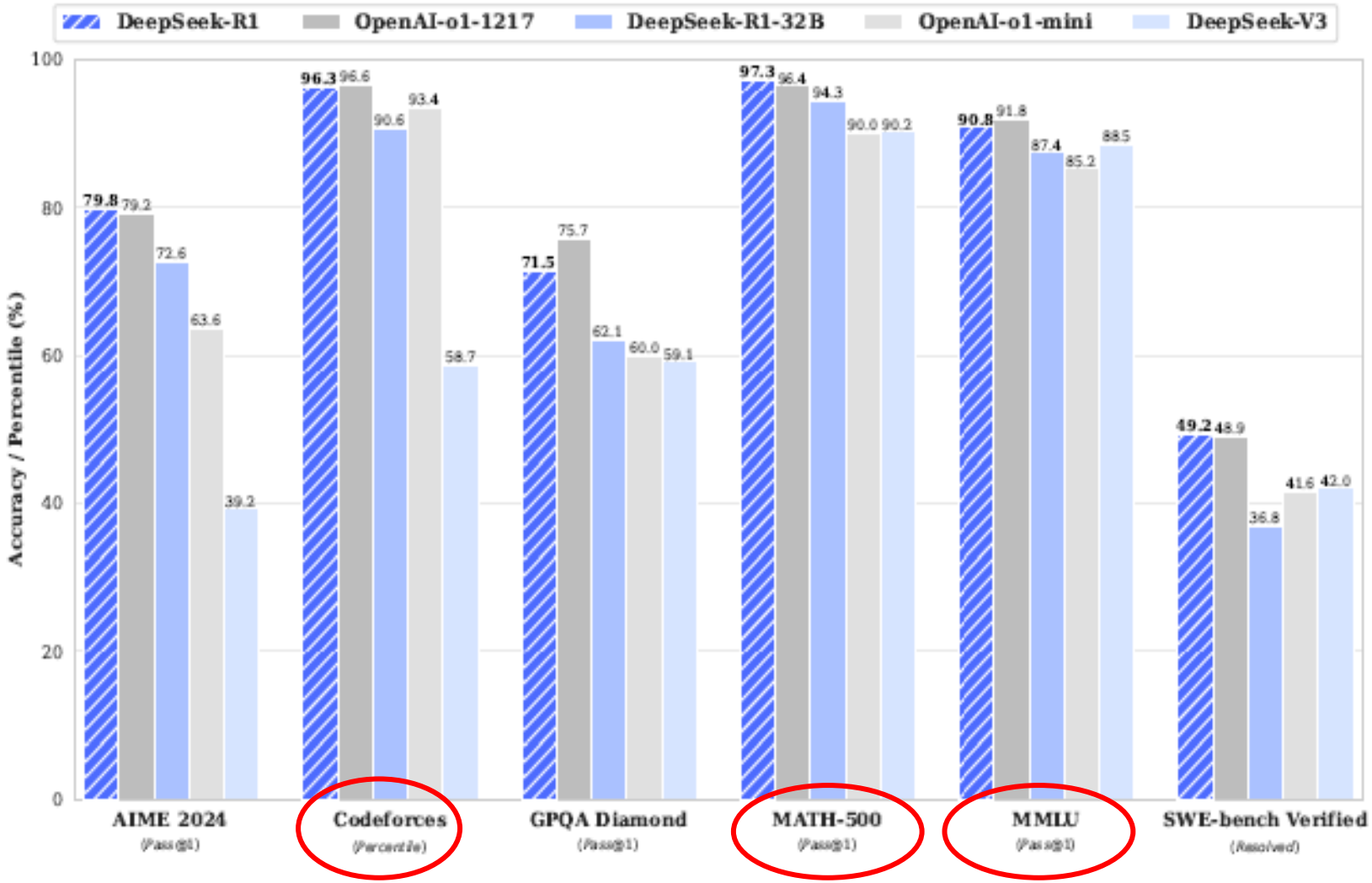


Figure 1 | Benchmark performance of DeepSeek-R1.

- AME2024:全米数学オリンピック
- Codeforces:競技プログラミング
- GPQA Diamond:大学院レベル自然科学
- MATH500:高校生向け数学コンテスト
- MMLU：多分野知識推論能力評価
- SWE-bench Varified:ソースコード修正課題

DeepSeek R1とは異なるGPT4.5のインパクト

■ DeepSeekのSLM（MoE）による論理的推論能力の高さ、そして「アハモーメント」が起こった。強化学習中に明確なプログラムなしで「反省」などの高度な推論を行った。

■ GPT4.5のインパクト

✓「教師なし学習」でパターン認識に優れている。ハルシネーションの大幅な低下がみられる。人間の思考能力の重要な特徴をなす直観的な回答をする。論理的推論能力はOpenAI O2、O3やDeepSeek R1より劣る。

✓ GPT 5 は直観能力と論理思考能力の相互作用を重視することになるだろう。より人間的思考に近づく！

→ Daniel Kahneman, *Thinking Fast and Slow*, 2011

[2] さらに高度な動きへ

Deep Mind-Google, O2 , GPT5(2026以降か）など

ノーベル化学賞2024　Alfa Fold3
デミス・ハサビス、ジョン・ジャンパー（Deep Mind, Google）

Google DeepMindのAlpha Fold3
May 8th 2024

■タンパク質の立体構造、DNA、mRNA、リガンドの相互作用や結合の予測

➡細胞の変化、発病の詳細な究明

➡因果論的アプローチから階層間相互作用と自己組織化のアプローチへ

➡創薬イノベーション、従来の生命科学・医学の抜本的なパラダイム・チェンジの端緒か

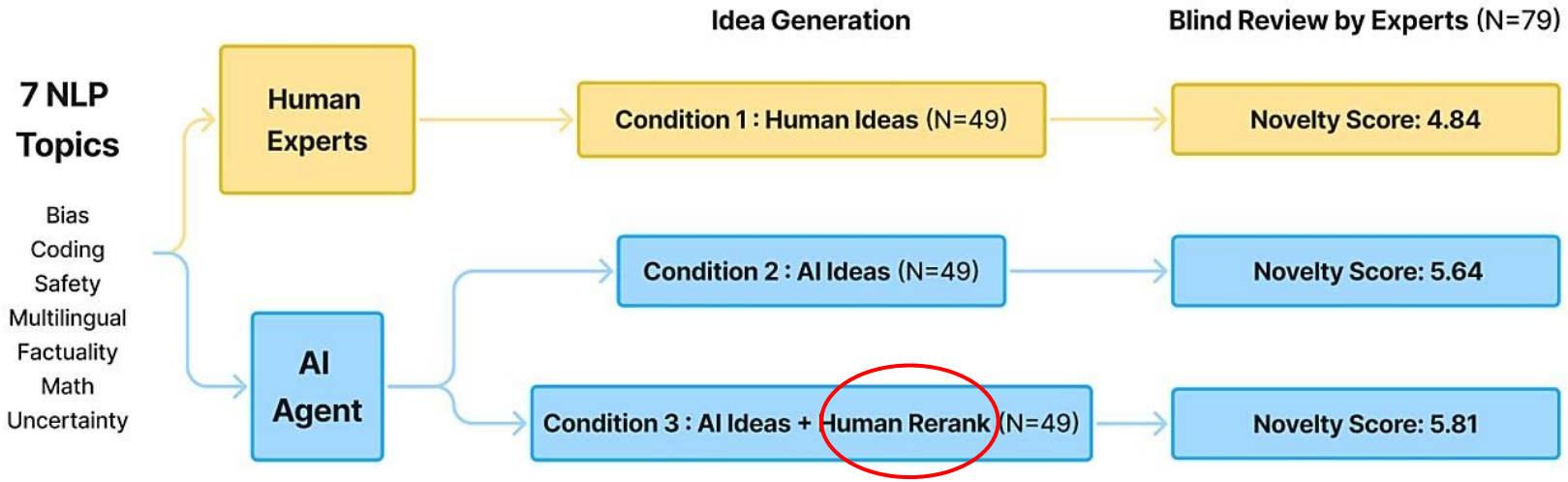
＊セコム科学技術振興財団 生命科学・医学分野専門家グループの検討においても、この動向を非常に重視している。マルチモーダルAIはこれからの最先端科学に必須。

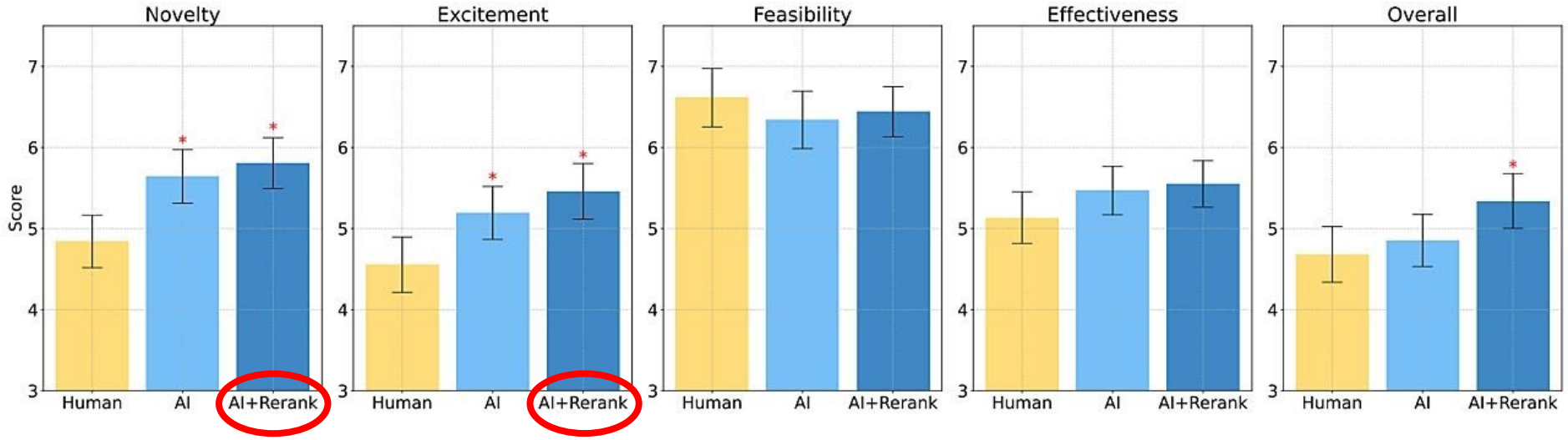
Gemini1.5が提案する新しい教育の一例

- Google 基調講演 (io.google)

- Googleは、LearnLMを活用したGoogle Classroomの新機能を試験的に導入し、教師の負担軽減を支援している。生成AI を活用して、授業計画のプロセスを簡素化し、教師が授業やコンテンツを生徒一人ひとりのニーズに合わせてカスタマイズできるようにする方法を構想し試行している。

AI Agent : LLMと人間のアイデア生成を比較した実験





Source： Chenglei Si, Diyi Yang, Tatsunori Hashimoto, Can LLMs Generate Novel Research Ideas? : A Large-Scale Human Study with 100+NLP Researchers, arXive, 6 Sep., 2024

個性とは？　そして創造性とは？

- 一般に人は誰でも、それぞれ異なった複数の活動領域をもち、それらの活動領域の情報をそれぞれ特有の仕方で関係づけることによって自らの活動を一貫性のあるものに統合している。

- 個性といわれるものは、既存の複数の活動領域の特性に規定され、さらに複数の活動領域から得られるさまざまな情報を関係づける既存の様式（文化）によって発現すると言ってよい。G.Simmel[1890]

- 須藤のいう創造性とは、主体的に複数の異なった事柄を関係づけ、さらには意図的に関係づける様式そのものを変化させる人間活動の特性である。須藤修[1995]

- The most exciting areas are in these fuzzy connections between disciplines where knowledge in one field answers questions in another field.

Prof. and Dr. Rita Colwell（元アメリカ合衆国NSF長官、東京大学大学院入学式祝辞、2013年4月12日）

・須藤修『複合的ネットーワーク社会』有斐閣、1995
・Georg Simmel, *Über sociale Differenzierung*, 1890

[3] 分裂する民主主義社会と生成AI

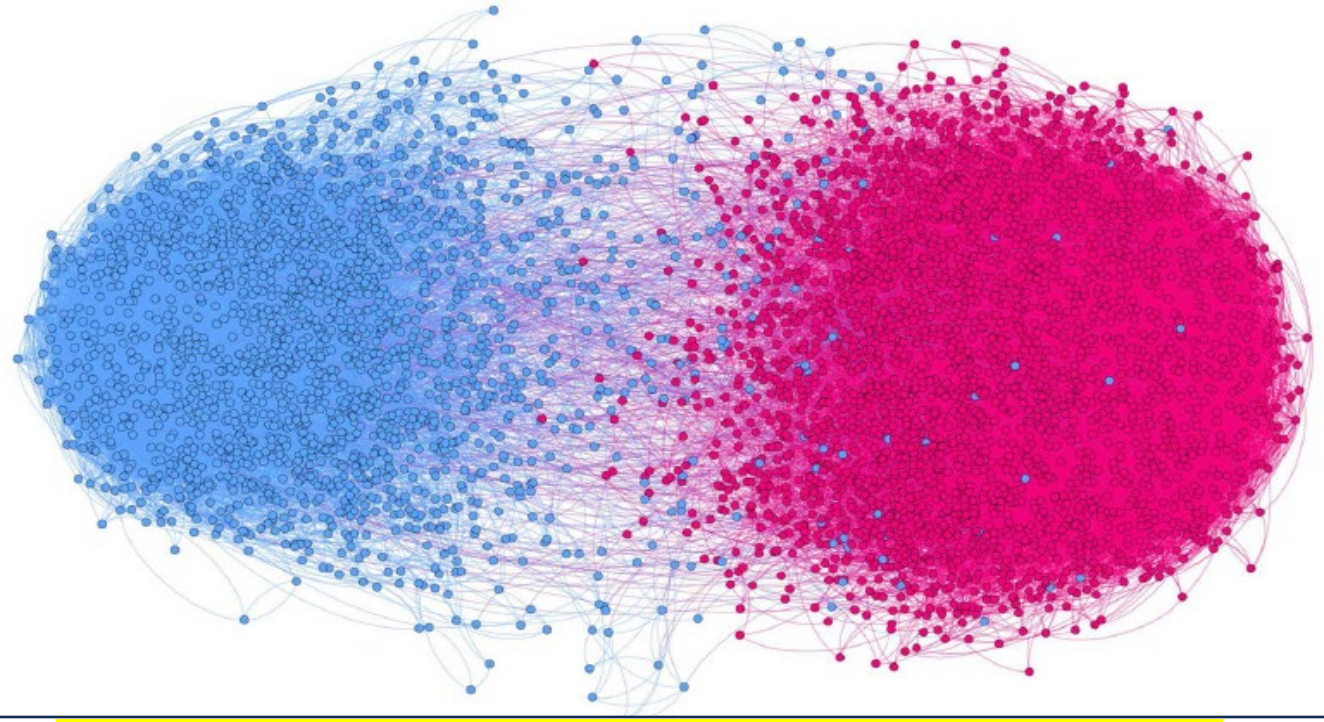
Nature Briefing　Nov. 6th 2024

「新しい世界に備える必要がある。ドナルド・トランプの反科学的な暴言と行動により、多くの科学者が、トランプが2期目の大統領に当選した今、科学への悪影響を覚悟していると語っている。彼らの懸念には、気候変動、公衆衛生、米国の民主主義のあり方などが含まれる。」

✓SNSの拡大とエコーチェンバーの増幅

✓民主主義のあるべき姿は？

Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, Filippo Menczer, Social influence and unfollowing accelerate the emergence of echo chambers, *Journal of Computational Social Science* (2021) 4



SNS：エコーチェンバーの発生

2010年のアメリカ合衆国中間選挙でのTwitterでの意見分布
青：民主党支持者、赤：共和党支持者

Habermas Machine by Deep Mind (1)

✓円滑な集団生活のためには、さまざまな意見をかけ合わせて合意に達する必要がある。しかし、正当性のある多くの意見が寄せられた場合、誰もが納得できる合意を得ることは困難である。そこで、イギリスAI安全研究所（AISI）の研究ディレクターである**クリストファー・サマーフィールド**は**Google DeepMind**とともに、AIが民主的な議論においてグループの合意形成を助けることができるかどうかを研究を行った。

✓研究チームはGoogle DeepMindが開発した大規模言語モデル「Chinchilla」をベースに「ハーバーマス・マシン」を作成。ちなみに、**「ハーバーマス・マシン」**という名前はドイツの社会哲学者ユルゲン・ハーバーマス氏（フランクフルト学派）にちなんで名付けられている。

✓ハーバーマス・マシンは、グループ内の個人の意見すべてを集約して、すべての人が受け入れられるような一連の**グループステートメント**を作成することが可能である。さらに、グループのメンバーは生成されたステートメントを評価して、システムへのフィードバックとトレーニングを実施できるほか、より完成度の高いステートメントを作成することも可能とされている。

Source: Michael Henry Tessler, Michiel A. Bakker et.al., AI can help humans find common ground in democratic deliberation, *Science*, 18 Oct. 2024, https://www.science.org/doi/10.1126/science.adq2852

Habermas Machine by Deep Mind (2)

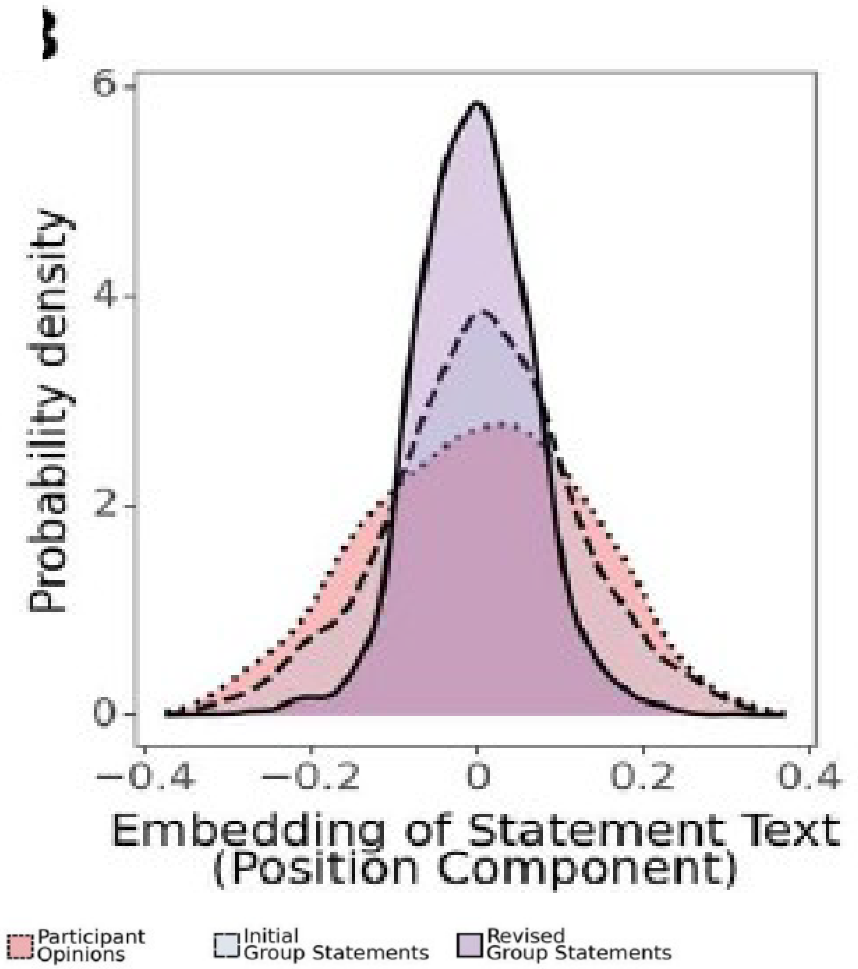
✓研究チームは439人のイギリス国民を集め、6人ずつ、計75グループに分けたうえで、イギリスの公共政策に関連する3つの議題(宗教教育、医療研究における動物実験など)について話し合い、各トピックについて個人的な意見を共有した後、包括的な意見を作成する実験を実施した。その際、各グループの参加者のうち1人が仲介役に指名され、グループ内の意見の取りまとめを行ったほか、同時にハーバーマス・マシンも取りまとめを実施した。その後、参加者には、人間の仲介役による取りまとめとハーバーマス・マシンが作成した取りまとめの両方が示され、どちらが優れているかを評価するよう求められた。

✓実験の結果、2つの取りまとめのうち、ハーバーマス・マシンが作成した取りまとめを支持したのは56%だったのに対して、人間の仲介役が作成した取りまとめを支持したのは44%で、参加者の過半数がAIが作成した取りまとめを評価していることが判明した。また、外部の審査員にも取りまとめの評価を依頼したところ、公平性や品質、明瞭さの点でハーバーマス・マシンが作成した取りまとめが高い評価を得たことが報告されている（N：5734名）。

✓また、研究チームはハーバーマス・マシンが参加者同士の議論の仲介を行うと、グループ内の合意のしやすさが仲介なしの場合よりも平均8%向上したことを伝えている。

Source: Michael Henry Tessler, Michiel A. Bakker et.al., AI can help humans find common ground in democratic deliberation, *Science*, 18 Oct. 2024, https://www.science.org/doi/10.1126/science.adq2852

Habermas Machine by Deep Mind (3)



RQ：Habermas Machineはすべての意見を平等に反映しているか？

意見から最初のグループ声明、修正されたグループ声明へとプロセスが進むにつれて、グループ声明がポジション軸に沿った意見間の妥協点を表すようになっている。

LLM生成AI Transformerの特性が出ている。

Source: Michael Henry Tessler, Michiel A. Bakker et.al., AI can help humans find common ground in democratic deliberation, *Science*, 18 Oct. 2024, https://www.science.org/doi/10.1126/science.adq2852

Habermas Machine by Deep Mind (4)

<MIT Tech Review 24 Oct. 2024>とのインタヴュー

■グーグル・ディープマインドの研究科学者、マイケル・ヘンリー・テスラーは語る。「この大規模言語モデルは、グループメンバー間で共有されているそれぞれの考えの中の重複領域を特定して提示するように訓練されている」。「人々を説得するように訓練されたのではなく、調停役となるように訓練されたのである」。

■＜須藤＞

✓調停役としてのAIの可能性拡大とその課題に関する研究が必要。

✓現時点では、調停役にはなりうるが、参加者を説得する役割を有する議長にはなれない。あるいは将来も議長にしてはいけないのだろうか？

✓大学はこのような課題に学融合的に取り組むべきではないか？**シニカルな姿勢はよくない**！

[4] 先進的AI社会とその高度なガバナンスを構想しよう！

Responsible AI（責任あるAI）という概念

✓プライバシーとデータガバナンス

✓透明性と説明可能性

✓セキュリティと安全性

✓公平性

- 以上4つの点が「責任あるAI」という概念のコアをなしている。今後は、とりわけ先進的AI（FM：基盤モデル、マルチモーダルAI）について計測可能性について熟考し、さまざまなツールを考案・改善し、人間のAIガバナンス能力を拡張すべきである。
- 特にヘルスケア、教育、金融、地球環境への影響は大きく、人間による先進AIのガバナンスの在り方に関する研究は極めて重要になる。

■ Multi-Stakeholdersの＜アゴラ＞としての大学

先進AIの社会展開の展望と課題
Responsible AIの在り方

■Multi-stakeholdersによる討議とは？

✓経済界の代表、市民社会の代表、行政機関代表、法曹界代表、教育機関代表、理工系研究者代表、生命系研究者代表、文系研究者代表、就労者代表など社会を構成する諸々の利害を代表する主体による討議

公民権・人権運動で偉大な功績を残したキング牧師の像（ワシントンD.C., 須藤修撮影2024）

これからの先進AIの展開について
検討すべき重要案件（須藤の意見）

- 言語理解の向上：例えば、最新の大規模言語モデル（LLM）は、学習データに基づいて説得力のあるテキストを生成することができるが、これらのモデルが生成した言語をどの程度理解しているかについては、かなりの議論の余地がある。➡ニューロ・シンボリックAIの研究

[Ref.] Amit Sheth, Kaushik Roy, Manas Gaur, "Neurosymbolic AI - Why, What, and How", *arXiv*, 1 May 2023

- AIの道徳的推論と人間の道徳的推論との整合性（アライメント）

Stanford Univ. の離散的測定研究など。

[Ref.] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn, Direct Preference Optimization : Your Language Model is Secretly a Reward Model, *arXiv*, 29 July 2024

- AGIに関する複数のシナリオの堅実かつ継続的な検討

OECD AI Futures(Aug. 2024), European AI Office(Aug. 2024), GPAI RAI-WG(2024)など。